Holzinger et al. BMC Bioinformatics 2014, 15(Suppl 6):I1
http://www.biomedcentral.com/1471-2105/15/S6/I1

BMC Bioinformatics

REVIEW Open Access

Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions

Andreas Holzinger[1,2*], Matthias Dehmer[3], Igor Jurisica[4,5]

Computers are incredibly fast, accurate, and stupid.

Human beings are incredibly slow, inaccurate, and brilliant.

Together they are powerful beyond imagination

(Einstein never said that [1]).

Background

The life sciences, biomedicine and health care are increasingly turning into a data intensive science [2-4]. Particularly in bioinformatics and computational biology we face not only increased volume and a diversity of highly complex, multi-dimensional and often weakly-structured and noisy data [5-8], but also the growing need for integrative analysis and modeling [9-14].

Due to the increasing trend towards personalized and precision medicine (P4 medicine: Predictive, Preventive, Participatory, Personalized [15]), biomedical data today results from various sources in different structural dimensions, ranging from the microscopic world, and in particular from the omics world (e.g., from genomics, proteomics, metabolomics, lipidomics, transcriptomics, epigenetics, microbiomics, fluxomics, phenomics, etc.) to the macroscopic world (e.g., disease spreading data of populations in public health informatics), see Figure 1[16]. Just for rapid orientation in terms of size: the Glucose molecule has a size of $900\,pm = 900 \times 10^{-12} m$ and the Carbon atom approx. $300\,pm$. A hepatitis virus is relatively large with $45\,nm = 45 \times 10^{-9} m$ and the X-Chromosome much bigger with $7\,\mu m = 7 \times 10^{-6} m$. We produce most of the "Big Data" in the omics world, we estimate many Terabytes ($1\,TB = 1 \times 10^{12}$ Byte = 1000 G Byte) of genomics data in each individual, consequently, the fusion of these with Petabytes of proteomics data for personalized medicine results in Exabytes of data ($1\,EB = 1 \times 10^{18}$ Byte). Last but not least, this "natural" data is then fused together with "produced" data, e.g., the unstructured information (text) in the patient records, wellness data, the data from physiological sensors, laboratory data etc. - these data are also rapidly increasing in size and complexity. Besides the problem of heterogeneous and distributed data, we are confronted with noisy, missing and inconsistent data. This leaves a large gap between the available "dirty" data [17] and the machinery to effectively process the data for the application purposes; moreover, the procedures of data integration and information extraction may themselves introduce errors and artifacts in the data [18].

Although, one may argue that "Big Data" is a buzz word, systematic and comprehensive exploration of all these data is often seen as the *fourth paradigm* in the investigation of nature - after empiricism, theory and computation [19], and provides a mechanism for data driven hypotheses generation, optimized experiment planning, precision medicine and evidence-based medicine.

The challenge is not only to extract meaningful information from this data, but to gain knowledge, to discover previously unknown insight, look for patterns, and to make sense of the data [20], [21]. Many different approaches, including statistical and graph theoretical methods, data mining, and machine learning methods, have been applied in the past - however with partly unsatisfactory success [22,23] especially in terms of performance [24].

The grand challenge is to make data useful to and useable by the end user [25]. Maybe, the key challenge is *interaction*, due to the fact that it is the human end user who possesses the problem solving intelligence [26], hence the ability to ask intelligent questions about the data. The problem in the life sciences is that (biomedical) data models are characterized by significant complexity [27], [28], making manual analysis by the end users difficult and often impossible [29]. At the same time, human

* Correspondence: a.holzinger@tugraz.at
[1]Research Unit Human-Computer Interaction, Austrian IBM Watson Think Group, Institute for Medical Informatics, Statistics & Documentation, Medical University Graz, Austria
Full list of author information is available at the end of the article

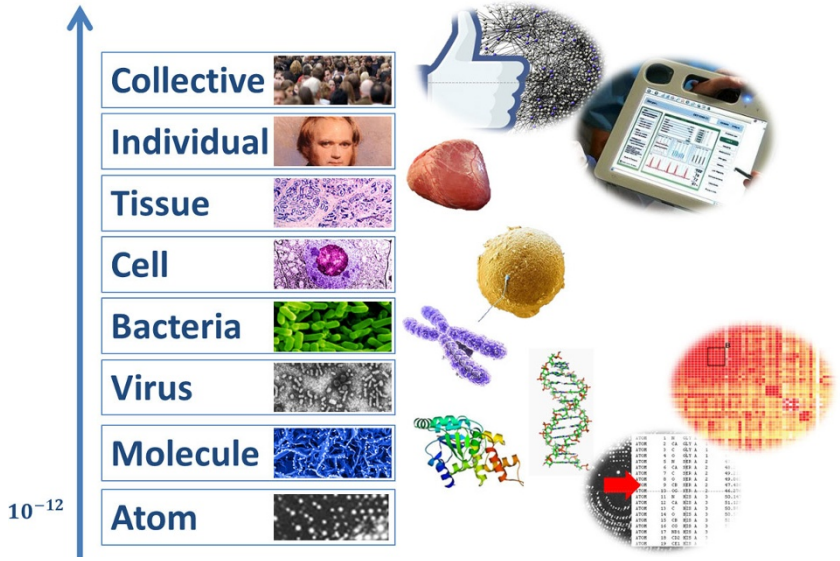

Figure 1 The trend towards personalized and molecular medicine brings together data from very different sources.

experts are able to solve complicated problems sometimes intuitively [30], [31], [32], e.g., often without being able to describe the exact rules or processes used during their analysis and problem solving.

Many advances in powerful computational tools [33], [34] in recent years have been developed by separate communities with different philosophies: Machine learning researchers tend to believe in the power of their statistical methods to identify relevant patterns [35] - mostly automatic, without human intervention [36]; however, the dangers of modelling artefacts grow when end user comprehension and control are diminished [37].

Additionally, mobile, ubiquitous computing and sensors, together with low cost storage, will accelerate this avalanche of data [38], and there will be a danger of drowning in data but starving for knowledge, as Herbert Simon pointed it out 40 years ago: *"A wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it"* [39].

Consequently, it is a grand challenge to work towards enabling effective human control over powerful machine intelligence by the integration and combination of machine learning methods and advanced visual analytics methods to support insight and decision making [28,40-44].

We envision effectively tackling these challenges by bringing together the best of two worlds: A synergistic combination of theories, methods and approaches from Human-Computer Interaction (HCI) and Knowledge Discovery from Data (KDD). Such approaches need a trans-disciplinary methodology. For example, the understanding of complex structures, such as regulatory networks, is a challenging objective and one that cannot be tackled within a single, isolated discipline [45]. Also, advances in network-based methods are enabled by novel applications. This relates to the exploration of methods and measures [46,47] to investigate global and local structural properties of complex networks or to study their interrelations [48-50]. While the relevant literature of the last decades has portrayed the definition of infinitely many network measures and methods as a relatively simply undertaking; overall, understanding this complex mathematical apparatus has turned out to be very complicated [51,52].

There is no doubt about the usefulness of such techniques in general. However, this branch of science somewhat failed to demonstrate the usefulness and interpretability of the underlying mathematical apparatus. In fact, while this development led to a vast amount of network measures/methods, exploring their structural interpretation and meaning has been often overlooked. This calls for generating more results to interpret the measures/methods more properly.

Knowledge Discovery process

The traditional method of turning data into knowledge relied on manual analysis and interpretation by a domain expert in order to find useful patterns in data for decision support. An early example from medical diagnostics includes the work by Reeder & Felson (1977) [53]. Today, far beyond pattern recognition, this process has been given a variety of names, including: data mining, knowledge extraction, information discovery, information harvesting,

data archaeology, and data pattern processing [54]. In the classic work by Fayyad et al. (1996), [55], this process is described by different steps starting from data selection, pre-processing, data transforming, data mining and interpretation. In this definition, Data Mining is actually a subset of Knowledge Discovery, and although the original notion was Knowledge Discovery in Databases (KDD), today, in order to emphasize that Data Mining is an important subset of the knowledge discovery process, the current most used notion is Knowledge Discovery and Data Mining (KDD). It is important to note that KDD can be seen as a *process* and encompasses the complete value added chain from the very physical side of data to the very human side of knowledge, the latter defined from a cognitive point of view: knowledge as a set of expectations [56]. We further extend the original definition by Fayyad et al. (1996) by *interaction* and include the human-into-the-loop. Interaction, communication and sensemaking are core topics in Human-Computer Interaction (HCI) [25,57-61], consequently, a novel approach is to combine HCI & KDD [8,44].

The central premise of HCI-KDD is to *enable* end users interactively to *find and characterize* previously unknown and potentially useful and usable information. It may be defined in the classical sense as the process of identifying novel data patterns, with the goal of understanding these patterns. The domain expert in Figure 2 possesses explicit domain knowledge and by enabling them to interactively explore the data sets, they may be able to identify, extract and understand useful information, to gain new, and previously unknown knowledge [21].

KDD historically builds on three fields: machine learning; databases and artificial intelligence to design and develop tools and frameworks that let the end users gain insight into the nature of massive data sets [54], [24], [62].

Future research directions

Figure 2 illustrates the complete knowledge discovery process, and we will use this "big picture" for the description of some problems and challenges - starting (in this Figure) from right to left - from the computer to the human - segmenting it into four large areas:

Area 1: Interactive data integration, data fusion and preselection of data sets

Many different biological species (humans, animals, bacteria, virus, plants, ...) deliver large amounts of data, together with the enormous complexity of medicine per se [42] and the limited computational power in comparison of the complexity of life (and the natural limitations of the Von-Neumann architecture) these pose a lot of problems, which can be divided into three categories:

- Heterogeneous data sources (need for data fusion);
- Complexity of the data (high-dimensionality);
- Noisy, uncertain data, dirty data, the discrepancy between data-information-knowledge (various definitions), Big data sets (when is data big? when manual handling of the data is impossible) [24].

In comparison to research systems, commercially available information systems have only limited data fusion capabilities, if any at all [63]. It is a huge challenge to integrate and fuse the biological data together with classical patient records, physiological data or medical image data [64], [65]. The issues are so big that there is an own conference series called "data integration in the life sciences" [66].

Area 2: Interactive sampling, cleansing, preprocessing, mapping

The problem of merging multiple data sets concerning common entities is frequently encountered in KDD, often called the Merge/Purge problem, it is difficult to solve both in scale and accuracy [67]. Cleansing data from impurities is an integral part of every data processing and has led to the development of a broad range of methods to enhance the accuracy and thereby the usability of existing data [68]. Many machine learning algorithms, for example, struggle with high-dimensional data. This has become well known as the curse of dimensionality [69]. A further issue is that most medical data is incomplete, with missing data values, inconsistent value naming conventions, etc. or requires the detection and removal of duplicate data entries [70] - so the central goal of data quality poses a number of problems and challenges [71], [72]. The quality of data finally, influences the quality of information [73].

Area 3: Interactive advanced data mining methods, pattern discovery

Many data mining methods are designed for collections of objects well-represented in rigid tabular formats. However, besides massive sets of unstructured information and non-standardized information (text) [74-76], we are increasingly confronted with large collections of interrelated objects whose natural representation is in point cloud data or typed graphs [77] (e.g., protein structures, protein interaction networks, etc.).

Advanced data mining approaches include:

1) graph-based data mining [78], [79], [80], [81],
2) entropy-based data mining [47,82], [83-85], and
3) topological data mining [86,87].

We emphasize that these approaches are interdisciplinary and complementary albeit having common goals, and

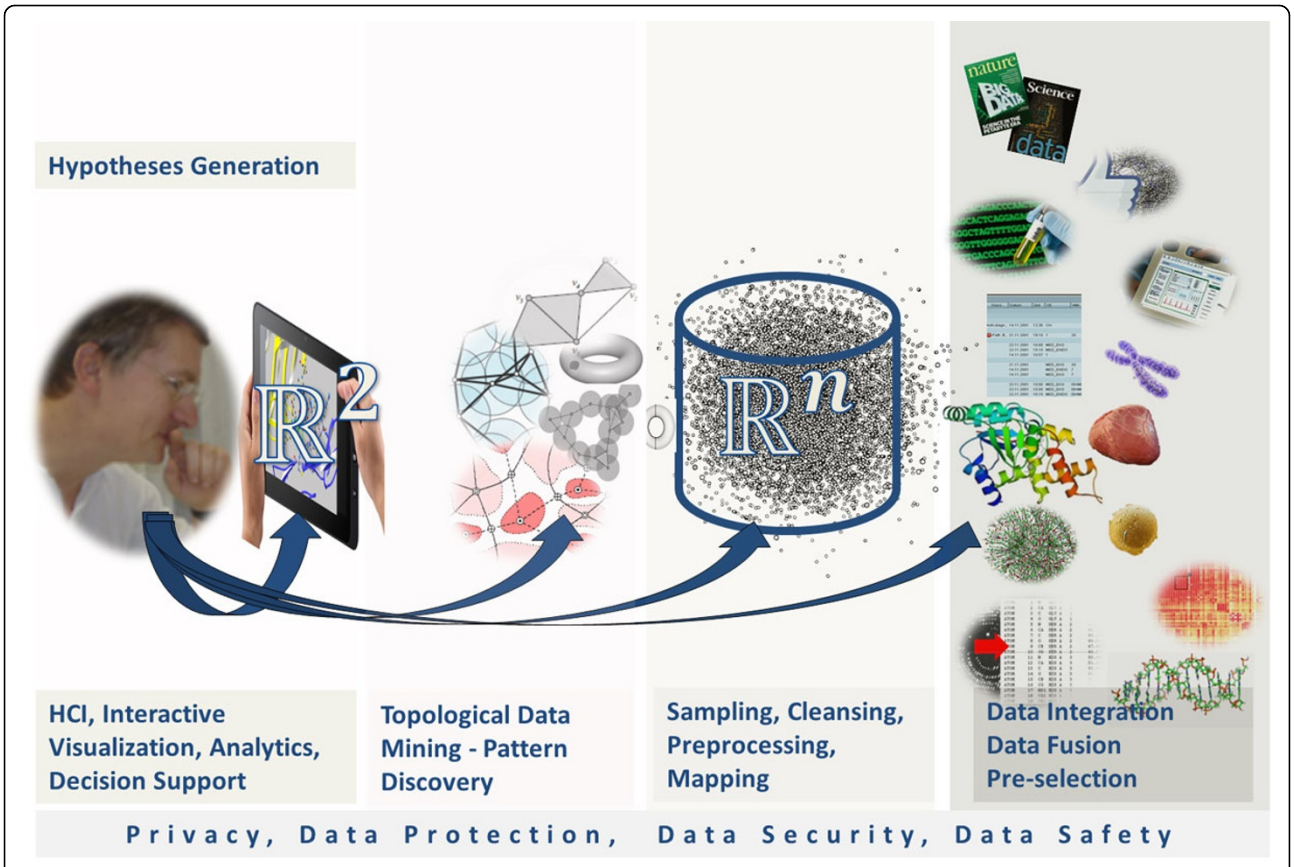

Figure 2 The knowledge discovery process in the life sciences.

have been proven useful to perform translational research, e.g., [47,82,84,85].

In particular, entropy-based graph analysis is based on using information theory and graph theory. Generally, information theory [88] relates to quantifying information and to investigating communication processes. To translate this concept to graph theory has been intricate. As a result, various graph entropies have been developed but the literature lacks exploring interrelations with other network measures. An example thereof can be found in [47]. Much future research is necessary in this area in the future.

Area 4: Interactive visualization, HCI, analytics, decision support

Finally, the results gained by the application of sophisticated algorithms in high dimensional spaces in area 3 must be mapped back to $\mathbb{R}^2$ because humans have difficulties in comprehending higher dimensional data.

We can say that, while our world is highly dimensional mathematically, we can only perceive lower dimensions. This leads to the definition of visualization *as the mapping from the higher into the lower dimensional space*, a process that always suffers the danger of modelling artefacts. Although Visualization is a mature field with a background of several decades, there are still a lot of challenging and open research issues, especially in the context of interactive data mining with application to the biomedical domain. A major issue is the absence of a complete toolset that supports all analysis tasks within a biomedical workflow, including the many steps of data preprocessing [89]. It is very interesting to note that although there are many sophisticated visualization techniques available [90-102], - these are rarely applied in routine applications, especially in business enterprise hospital information systems, where such approaches really could bring benefits to the professionals. An extremely important issue is the limited time, e.g., in average a medical doctor in a public hospital has only five minutes to make a decision [103,104]; This strongly calls for interactive tools. Naive visualization attempts are often ineffective or even actively misleading, due to the fact that the development of effective visualizations is a complex process and requiring a basic understanding of human information-processing and a solid grounding in the existing body of work in the visualization community [105-107].

Horizontal area: Privacy, data protection, data security, data safety

Whenever we deal with biomedical data issues of privacy, data protection, data security and data safety and the fair use of data are of paramount importance [108], including data accessibility, temporal limits, legal restrictions (such as in situations where copyright or patents may be relevant), confidentiality and data provenance. We face a range of research challenges in developing data mining methods to properly handle these complex restrictions.

Additional aspects to consider

Some additional aspects to consider include:

Cross-disciplinary cooperation with domain experts

Building a project consortium comprising of experts with complementary expertise but common interests is a success factor in each project. Bringing together domain experts from diverse areas in a cross-disciplinary manner is a challenge to stimulate fresh ideas and encouraging multi-disciplinary work [109]. For example, the application of principles from HCI to data-driven projects in biomedical contexts has been lacking and has been receiving increasing attention in recent years [59], [110]. In the life sciences domain, experts are both data producers and end users at the same time, knowledge engineers and analysts help to organize, integrate, visualize, analyze and evaluate the data. For example, in "systems biology" intertwining these two may lead to improving both the models and the experimental results. In such complex domains as in biomedicine, we need experts who understand the domain, the problem, and the data sets, hence the context [111].

Interpretability

As we broaden workflows for data mining, we have to expand metrics used to evaluate our results. It is no longer sufficient to focus on performance metrics, such as ROC [112], accuracy, precision and recall (although precision and recall still are *the* measures in data mining [113]), one must also consider how non-functional requirements are satisfied, such as interpretability. In the biomedical domain, where it is necessary to explain or justify the results of a decision, data mining alone is definitely irrelevant: It is necessary to produce results that are explainable to others. In a SIAM conference in 2007 an interesting panel was held, where the panelists including Christos Faloutsos (Carnegie Mellon University), Jerry Friedman (Stanford University), Ajay Royyuru (IBM Research), and Mehran Sahami (Google Research), together with the moderator Haym Hirsh (Rutgers University), formulated a couple of interesting questions, which are very relevant up to the present [23], for example: How can we quantitatively and qualitatively measure interpretability? Similar to the concepts of interest or beauty [114], interpretability is in the eye of the beholder and possibly dependent on the previous knowledge and the level of expertise of the decision maker [115], consequently, we need adaptive tools to satisfy both novices and experts.

Computing resources

As our computing machinery evolves, from large mainframe servers to multi-core CPU/GPU clusters we need to optimize data mining algorithms, processes and workflows to best fit the environment. The potential of so-called On-Demand Hardware along with the Software as a Service (SAAS) paradigm [116] can no longer be denied, and there are several examples yet of Cloud Computing approaches, e.g. in drug discovery research, medical imaging and applications for doctors in rural areas [117-119]. However, much data in biomedicine and healthcare has strict privacy requirements and therefore privacy, security safety and data protection issues are of enormous importance with such future approaches. Major internet companies offer already such services for data-intensive computing and a similar strategy led to the developing of large computing grids for massive data analysis, such as IBM's World Community Grid (http://www.worldcommunitygrid.org), [120].

Benchmarking against gold-standards

To measure the quality of data mining approaches, the production of benchmarks it very important. These data sets can be used as so-called gold-standards (e.g., [121-123], which allow us to compare results across competing methods and are thus important for information quality issues [124,125].

Reproducibility

A big general issue among our modern research communities is that rarely one can reproduce the results of other researchers. Often it is not possible to verify and to replicate experiments, which is the case for example in classical non-computing experimental sciences [126]. One of the major issues is "sloppiness in data handling" and the resulting exponentially growing retraction of papers [127]. So, a mega challenge is in ensuring that results can be replicated from other groups at other places.

Embedded data mining

Whilst existing research has shown the value of data-driven science, we need to further integrate knowledge discovery and visualization pipelines into biological and biomedical and especially clinical workflows to take full advantage of their potential [23].

Complexity of data analysis methods

Deciding which method is the most suitable for solving a particular data analysis problem is often critical as the interdependencies make the selection non-linear [128]. Hence to perform data analysis efficiently, a deep understanding of the underlying mathematical apparatus is necessary.

Conclusion

We are just at the beginning of a turning point towards data intensive life sciences, which entails many challenges and future research directions. Within this overview we have highlighted only a few issues. Summarizing, we may say that the grand challenge is in building frameworks for enabling domain experts to interactively deal with their data sets in order to "ask questions" about the data, for example: "Show me similarities/differences/anomalies of data set × and data set Y", hence the discovery of novel, previously unknown patterns in complex data. Which mathematical framework should we use? One challenge is that such a framework must be usable for domain experts without prior training in mathematics or computational sciences. We need machine intelligence to deal with the flood of data, but at the same time we must acknowledge that humans possess certain problem solving and cognition abilities, which are far beyond computation. A possible solution is in the cross-disciplinary combination of aspects of the better of two worlds: Human-Computer Interaction (HCI) and Knowledge Discovery from Data (KDD). A proverb attributed perhaps incorrectly to Albert Einstein illustrates this perfectly: "Computers are incredibly fast, accurate, but stupid. Humans are incredibly slow, inaccurate, but brilliant. Together they may be powerful beyond imagination".

Competing interests

All authors declare that they have no competing interests.

Authors' information

Andreas Holzinger is head of the Research Unit Human-Computer Interaction, Institute for Medical Informatics at the Medical University Graz, Lead at the HCI-KDD network, head of the first Austrian IBM Watson Think Group, Associate Professor of Applied Informatics at the Faculty of Computer Science, Institute of Information Systems and Computer Media and Lecturer at the Institute of Genomics and Bioinformatics at Graz University of Technology. He serves as consultant for the Canadian, Swiss, French and Dutch Government, for the German Excellence Initiative and as national expert in the European Commission (Lisbon Delegate 2000). Andreas, born 1963, started as an apprentice in IT in 1978; while working as an industrial engineer, he resumed a parallel second-chance education, finished his PhD in Cognitive Science in 1997 and completed his second doctorate (Habilitation) in Computer Science in 2003. Since 1999 participation in leading positions in 30+ R&D multi-national projects, budget 3+ MEUR; 300+ publications, >4000+ citations. Andreas was Visiting Professor in Berlin, Innsbruck, Vienna, London, and Aachen. He is passionate on bringing together Human-Computer Interaction (HCI) and Knowledge Discovery/Data Mining (KDD), with the goal of supporting human intelligence with machine intelligence - to discover new, previously unknown insights into complex biomedical data. http://www.hci4all.at

Matthias Dehmer is currently head of the Division for Bioinformatics and Translational Research at UMIT, Austria and a professor of discrete mathematics. He studied mathematics and computer science at the University of Siegen (Germany) where he graduated in 1998. Between 1998 and 2002, he held positions as a mathematical researcher and a business consultant in industry. He joined in 2002 the Department of Computer Science at Darmstadt University of Technology and obtained a PhD in computer science. From 2005 to 2008, he held several research positions at the University of Rostock (Germany), Vienna Bio Center (Austria), Vienna Technical University (Austria) and University of Coimbra (Portugal). Finally, he obtained his habilitation in applied discrete mathematics from the Vienna University of Technology. He has been the head of the Division for Bioinformatics and Translational Research at UMIT, Austria. He has published over 160 publications in applied mathematics and computer science. Moreover, he is an editor of the book series "Quantitative and Network Biology", Wiley-VCH. He organized and co-organized several international scientific conferences and workshops in USA. Also, he recently got a member of the editorial board of Scientific Reports (Nature) and PLoS ONE. http://www.dehmer.org

Igor Jurisica is Tier I Canada Research Chair in Integrative Cancer Informatics, is a Senior Scientist at Princess Margaret Cancer Centre, Professor at the University of Toronto and Visiting Scientists at IBM's Centre for Advanced Studies. He is also an Adjunct Professor at the School of Computing, Department of Pathology and Molecular Medicine Queen's U and Department of Computer Science and Engineering at York University. Igor's research focuses on integrative computational biology and the representation, analysis and visualization of high-dimensional data to identify prognostic/predictive signatures, drug mechanism of action and in silico repositioning of drugs. Interests include comparative analysis for mining different integrated data sets (e.g., protein-protein interactions, high-dimensional cancer data, and high-throughput screens for protein crystallization). http://www.cs.toronto.edu/~juris.

Declarations

Publication for this article has been funded by the Research Unit hci4all.at
This article has been published as part of *BMC Bioinformatics* Volume 15 Supplement 6, 2014: Knowledge Discovery and Interactive Data Mining in Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/15/S6.

Authors' details

[1]Research Unit Human-Computer Interaction, Austrian IBM Watson Think Group, Institute for Medical Informatics, Statistics & Documentation, Medical University Graz, Austria. [2]Institute of Information Systems and Computer Media, Graz University of Technology, Austria. [3]Institute for Bioinformatics and Translational Research, UMIT Tyrol, Austria. [4]Departments of Medical Biophysics and Computer Science, University of Toronto, Ontario, Canada. [5]Princess Margaret Cancer Centre and Techna Institute for the Advancement of Technology for Health, University Health Network, IBM Life Sciences Discovery Centre, Ontario, Canada.

Published: 16 May 2014

References

1. **Einstein never said that.** [http://www.benshoemate.com/2008/11/30/einstein-never-said-that].
2. Ranganathan S, Schonbach C, Kelso J, Rost B, Nathan S, Tan T: **Towards big data science in the decade ahead from ten years of InCoB and the 1st ISCB-Asia Joint Conference.** *BMC Bioinformatics* 2011, **12(Suppl 13)**:S1.
3. Dhar V: **Data science and prediction.** *Communication of the ACM* 2013, **56(12)**:64-73.
4. Kolker E, Özdemir V, Martens L, Hancock W, Anderson G, Anderson N, Aynacioglu S, Baranova A, Campagna SR, Chen R: **Toward more transparent and reproducible omics studies through a common metadata checklist and data publications.** *OMICS: A Journal of Integrative Biology* 2014, **18(1)**:10-14.
5. Morik K, Brockhausen P, Joachims T: **Combining statistical learning with a knowledge-based approach-a case study in intensive care monitoring.** *ICML* 1999, **99**:268-277.

6. Sultan M, Wigle DA, Cumbaa C, Maziarz M, Glasgow J, Tsao M, Jurisica I: **Binary tree-structured vector quantization approach to clustering and visualizing microarray data.** *Bioinformatics* 2002, **18(suppl 1)**:S111-S119.
7. Holzinger A: **Weakly Structured Data in Health-Informatics: The Challenge for Human-Computer Interaction.** In *Proceedings of INTERACT 2011 Workshop: Promoting and supporting healthy living by design; Lisbon (Portugal).* IFIP;Baghaei N, Baxter G, Dow L, Kimani S 2011:5-7.
8. Holzinger A: **On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data - Challenges in Human-Computer Interaction & Biomedical Informatics.** *DATA 2012* Rome, Italy: INSTICC; 2012, 9-20.
9. Olshen AB, Hsieh AC, Stumpf CR, Olshen RA, Ruggero D, Taylor BS: **Assessing gene-level translational control from ribosome profiling.** *Bioinformatics* 2013, **29(23)**:2995-3002.
10. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22(13)**:1658-1659.
11. Pržulj N, Wigle D, Jurisica I: **Functional topology in a network of protein interactions.** *Bioinformatics* 2004, **20(3)**:340-348.
12. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *BMC Bioinformatics* 2010, **11.**
13. Kiberstis PA: **All Eyes on Epigenetics.** *Science* 2012, **335(6069)**:637.
14. Barrera J, Cesar-Jr RM, Ferreira JE, Gubitoso MD: **An environment for knowledge discovery in biology.** *Computers in Biology and Medicine* 2004, **34(5)**:427-447.
15. Hood L, Friend SH: **Predictive, personalized, preventive, participatory (P4) cancer medicine.** *Nature Reviews Clinical Oncology* 2011, **8(3)**:184-187.
16. Holzinger A: **Biomedical Informatics: Discovering Knowledge in Big Data.** New York: Springer; 2014.
17. Kim W, Choi B-J, Hong E-K, Kim S-K, Lee D: **A taxonomy of dirty data.** *Data Min Knowl Discov* 2003, **7(1)**:81-99.
18. Ouzzani M, Papotti P, Rahm E: **Editorial: Introduction to the special issue on data quality.** *Information Systems* 2013, **38(6)**:885-886.
19. Bell G, Hey T, Szalay A: **Beyond the data deluge.** *Science* 2009, **323(5919)**:1297-1298.
20. Holzinger A, Stocker C, Bruschi M, Auinger A, Silva H, Fred A: **On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data.** In *Active Media Technologies AMT 2012, LNCS 7669.* Springer;R. Huang eaE. Macau 2012:646-657.
21. Holzinger A, Scherer R, Seeber M, Wagner J, Müller-Putz G: **Computational Sensemaking on Examples of Knowledge Discovery from Neuroscience Data: Towards Enhancing Stroke Rehabilitation.** In *Information Technology in Bio- and Medical Informatics, Lecture Notes in Computer Science, LNCS 7451.*. New York,Springer;Böhm C, Khuri S, Lhotská L, Renda M. Heidelberg 2012:166-168.
22. Raymer ML, Doom TE, Kuhn LA, Punch WF: **Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/ evolutionary algorithm.** *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics* 2003, **33(5)**:802-813.
23. Hirsh H: **Data mining research: Current status and future opportunities.** *Statistical Analysis and Data Mining* 2008, **1(2)**:104-107.
24. Jurisica I, Mylopoulos J, Glasgow J, Shapiro H, Casper RF: **Case-based reasoning in IVF: prediction and knowledge mining.** *Artificial intelligence in medicine* 1998, **12(1)**:1-24.
25. Holzinger A, Thimbleby H, Beale R: **Human-Computer Interaction for Medicine and Health Care (HCI4MED): Towards making Information usable.** *International Journal of Human-Computer Studies (IJHCS)* 2010, **28(6)**:325-327.
26. Simon HA: **Studying Human Intelligence by Creating Artificial Intelligence.** *American Scientist* 1981, **69(3)**:300-309.
27. Akil H, Martone ME, Van Essen DC: **Challenges and opportunities in mining neuroscience data.** *Science* 2011, **331(6018)**:708-712.
28. Holzinger A: **Biomedical Informatics: Computational Sciences meets Life Sciences.** Norderstedt: BoD; 2012.
29. Dugas M, Schmidt K: **Medizinische Informatik und Bioinformatik.** Berlin, Heidelberg: Springer; 2003.
30. Polanyi M: **Personal Knowledge: Towards a Post-Critical Philosophy.** Nature Publishing Group; 1974.
31. Popper KR: *Alles Leben ist Problemlösen* München, Zürich: Piper; 1996.
32. D'Negri CE, De Vito EL: **Making it possible to measure knowledge, experience and intuition in diagnosing lung injury severity: a fuzzy logic vision based on the Murray score.** *BMC Med Inform Decis Mak* 2010, 10.
33. Kruse R, Borgelt C, Klawonn F, Moewes C, Steinbrecher M, Held P: **Computational Intelligence: A Methodological Introduction.** Heidelberg, New York: Springer; 2013.
34. Holzinger A, Zupan M: **KNODWAT: A scientific framework application for testing knowledge discovery methods for the biomedical domain.** *BMC Bioinformatics* 2013, **14(1)**:191.
35. Zhou J, Lamichhane S, Sterne G, Ye B, Peng H: **BIOCAT: a pattern recognition platform for customizable biological image classification and annotation.** *BMC Bioinformatics* 2013, **14(1)**:291.
36. Gao H, Siu WC, Hou CH: **Improved techniques for automatic image segmentation.** *Ieee Transactions on Circuits and Systems for Video Technology* 2001, **11(12)**:1273-1280.
37. Shneiderman B: **Inventing Discovery Tools: Combining Information Visualization with Data Mining.** *Information Visualization* 2002, **1(1)**:5-12.
38. Butler D: **2020 computing: Everything, everywhere.** *Nature* 2006, **440(7083)**:402-405.
39. Simon HA: **Designing Organizations for an Information-Rich World.** In *Computers, Communication, and the Public Interest.* The Johns Hopkins Press; Greenberger M. Baltimore (MD) 1971:37-72.
40. Holzinger A: **Interacting with Information: Challenges in Human-Computer Interaction and Information Retrieval (HCI-IR).** *IADIS Multiconference on Computer Science and Information Systems (MCCSIS), Interfaces and Human-Computer Interaction* Rome: IADIS; 2011, 13-17.
41. Shortliffe EH: **Biomedical Informatics: Defining the Science and its Role in Health Professional Education.** In *Information Quality in e-Health Lecture Notes in Computer Science LNCS 7058.* New York: Springer;Holzinger A, Simonic K-M. Heidelberg 2011:711-714.
42. Patel VL, Kahol K, Buchman T: **Biomedical Complexity and Error.** *J Biomed Inform* 2011, **44(3)**:387-389.
43. Bloice M, Simonic K-M, Kreuzthaler M, Holzinger A: **Development of an Interactive Application for Learning Medical Procedures and Clinical Decision Making.** In *Information Quality in e-Health (Lecture Notes in Computer Science LNCS 7058). Volume 7058.* Berlin, Heidelberg, New York: Springer;Holzinger A, Simonic K-M 2011:211-224.
44. Holzinger A: **Human-Computer Interaction & Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together?** In *Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127.*. Heidelberg, Berlin, New York: Springer;Alfredo Cuzzocrea CK, Dimitris E. Simos, Edgar Weippl, Lida Xu 2013:319-328.
45. Beslon G, Parsons DP, Peña J-M, Rigotti C, Sanchez-Dehesa Y: **From digital genetics to knowledge discovery: Perspectives in genetic network understanding.** *Intelligent Data Analysis* 2010, **14(2)**:173-191.
46. Dehmer M: **Structural analysis of complex networks.** Birkhäuser Boston; 2010.
47. Dehmer M, Mowshowitz A: **A history of graph entropy measures.** *Inf Sci* 2011, **181(1)**:57-78.
48. Pržulj N, Corneil DG, Jurisica I: **Modeling interactome: scale-free or geometric?** *Bioinformatics* 2004, **20(18)**:3508-3515.
49. Dehmer M, Borgert S, Bonchev D: **Information inequalities for graphs.** In *Symmetry: Culture and Science Symmetry in Nanostructures* M Diudea 2008, **19(Special)**:269-284.
50. Dehmer M, Varmuza K, Borgert S, Emmert-Streib F: **On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures.** *Journal of chemical information and modeling* 2009, **49(7)**:1655-1663.
51. Randic M: **On molecular identification numbers.** *Journal of Chemical Information and Computer Sciences* 1984, **24(3)**:164-175.
52. Todeschini R, Consonni V: *Handbook of molecular descriptors* Hoboken (NJ): Wiley; 2008.
53. Reeder MM, Felson B: *Gamuts in radiology: comprehensive lists of roentgen differential diagnosis* Pergamon Press; 1977.
54. Fayyad U, Piatetsky-Shapiro G, Smyth P: **The KDD process for extracting useful knowledge from volumes of data.** *Communications of the ACM* 1996, **39(11)**:27-34.
55. Fayyad U, Piatetsky-Shapiro G, Smyth P: **From data mining to knowledge discovery in databases.** *Ai Magazine* 1996, **17(3)**:37-54.

56. Boisot M, Canals A: **Data, information and knowledge: have we got it right?** *Journal of Evolutionary Economics* 2004, **14(1)**:43-67.
57. Nake F, Grabowski S: **Human-Computer Interaction viewed as Pseudo-Communication.** *Knowledge-Based Systems* 2001, **14(8)**:441-447.
58. Holzinger A, Ackerl S, Searle G, Sorantin E: **Speech Recognition in daily Hospital practice: Human-Computer Interaction Lessons learned.** *CEMVRC (Central European Multimedia and Virtual Reality Conference): 2004; Vezprém (Hungary)* University of Pannonia Press;125-134.
59. Blandford A, Attfield S: **Interacting with Information.** *Synthesis Lectures on Human-Centered Informatics* 2010, **3(1)**:1-99.
60. Blandford A, Faisal S, Attfield S: **Conceptual design for sensemaking.** *Handbook of Human Centric Visualization* New York: Springer; 2014, 253-283.
61. Holzinger A, Searle G, Auinger A, Ziefle M: **Informatics as Semiotics Engineering: Lessons Learned from Design, Development and Evaluation of Ambient Assisted Living Applications for Elderly People.** In *Universal Access in Human-Computer Interaction Context Diversity, Lecture Notes in Computer Science, LNCS 6767.* Berlin, Heidelberg: Springer;Stephanidis C 2011:183-192.
62. Yildirim P, Ekmekci I, Holzinger A: **On Knowledge Discovery in Open Medical Data on the Example of the FDA Drug Adverse Event Reporting System for Alendronate (Fosamax).** In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Lecture Notes in Computer Science, LNCS 7947.* Berlin, Heidelberg: Springer;Holzinger A, Pasi G 2013:195-206.
63. Bleiholder J, Naumann F: **Data fusion.** *ACM Computing Surveys (CSUR)* 2008, **41(1)**:1.
64. Wiltgen M, Holzinger A, Groell R, Wolf G, Habermann W: **Usability of Image fusion: optimal opacification of vessels and squamous cell carcinoma in CT scans.** *Springer Elektrotechnik & Informationstechnik, e&i* 2006, **123(4)**:156-162.
65. Viceconti M, Taddei F, Montanari L, Testi D, Leardini A, Clapworthy G, Jan SV: **Multimod data manager: A tool for data fusion.** *Computer Methods and Programs in Biomedicine* 2007, **87(2)**:148-159.
66. Baker CJ, Butler G, Jurisica I: **Data Integration in the Life Sciences: 9th International Conference, DILS 2013, Montreal, Canada, July 11-12, 2013, Proceedings.** Springer Publishing Company, Incorporated; 2013.
67. Hernández MA, Stolfo SJ: **Real-world data is dirty: Data cleansing and the merge/purge problem.** *Data Min Knowl Discov* 1998, **2(1)**:9-37.
68. Müller H, Freytag J-C: **Problems, methods, and challenges in comprehensive data cleansing.** Professoren des Inst. Für Informatik; 2005.
69. Catchpoole DR, Kennedy P, Skillicorn DB, Simoff S: **The Curse of Dimensionality: A Blessing to Personalized Medicine.** *J Clin Oncol* 2010, **28(34)**:E723-E724.
70. Lee ML, Lu H, Ling TW, Ko YT: **Cleansing data for mining and warehousing.** *Database and Expert Systems Applications: 1999* Springer;751-760.
71. Elloumi M, Zomaya AY: **Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data.** John Wiley & Sons; 2013**23.**
72. Jarke M, Jeusfeld MA, Quix C, Vassiliadis P: **Architecture and Quality in Data Warehouses.** *Seminal Contributions to Information Systems Engineering.* Springer; 2013, 161-181.
73. Holzinger A, Simonic K-M: **Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058.** Heidelberg, Berlin, New York: Springer; 2011.
74. Kreuzthaler M, Bloice M, Simonic K-M, Holzinger A: **Navigating through Very Large Sets of Medical Records: An Information Retrieval Evaluation Architecture for Non-standardized Text.** In *Information Quality in e-Health, Lecture Notes in Computer Science, LNCS 7058.* Springer Berlin Heidelberg; Holzinger A, Simonic K-M 2011:455-470.
75. Holzinger A, Simonic KM, Yildirim P: **Disease-Disease Relationships for Rheumatic Diseases: Web-Based Biomedical Textmining an Knowledge Discovery to Assist Medical Decision Making.** *IEEE 36th Annual Computer Software and Applications Conference (COMPSAC): 16-20 July 2012 2012; Izmir, Turkey* 573-580.
76. Petz G, Karpowicz M, Fürschuß H, Auinger A, Winkler S, Schaller S, Holzinger A: **On Text Preprocessing for Opinion Mining Outside of Laboratory Environments.** In *Active Media Technology, Lecture Notes in Computer Science, LNCS 7669.*. Berlin Heidelberg: Springer;Huang R, Ghorbani A, Pasi G, Yamaguchi T, Yen N, Jin B 2012:618-629.
77. Jurack S, Taentzer G: **A component concept for typed graphs with inheritance and containment structures.** *Graph Transformations* Springer; 2010, 187-202.
78. Cook DJ, Holder LB: **Graph-based data mining.** *IEEE Intell Syst Appl* 2000, **15(2)**:32-41.
79. Fischer I, Meinl T: **Graph based molecular data mining - An overview.** *2004 IEEE International Conference on Systems, Man & Cybernetics.* New York: IEEE; 2004, 4578-4582.
80. Wang F, Jin R, Agrawal G, Piontkivska H: **Graph and topological structure mining on scientific articles.** In *Proceedings of the 7th Ieee International Symposium on Bioinformatics and Bioengineering, Vols I and Ii; New York.* IEEE;Yang MQ, Zhu MM, Zhang Y, Arabnia HR, Deng Y, Bourbakis N 2007:1318-1322.
81. Shelokar P, Quirin A, Cordon O: **A multiobjective evolutionary programming framework for graph-based data mining.** *Inf Sci* 2013, **237**:118-136.
82. Koslicki D: **Topological entropy of DNA sequences.** *Bioinformatics* 2011, **27(8)**:1061-1067.
83. Holzinger A, Stocker C, Peischl B, Simonic K-M: **On Using Entropy for Enhancing Handwriting Preprocessing.** *Entropy* 2012, **14(11)**:2324-2350.
84. Holzinger A, Stocker C, Bruschi M, Auinger A, Silva H, Gamboa H, Fred A: **On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data.** In *Active Media Technology, Lecture Notes in Computer Science, LNCS 7669.*. Berlin Heidelberg: Springer; Huang R, Ghorbani A, Pasi G, Yamaguchi T, Yen N, Jin B 2012:646-657.
85. Holzinger A, Ofner B, Stocker C, Valdez AC, Schaar AK, Ziefle M, Dehmer M: **On Graph Entropy Measures for Knowledge Discovery from Publication Network Data.** In *Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127.*. Heidelberg, Berlin: Springer;Cuzzocrea A, Kittl C, Simos DE, Weippl E, Xu L 2013:354-362.
86. Singh G, Mémoli F, Carlsson G: **Topological methods for the analysis of high dimensional data sets and 3D object recognition.** *Eurographics Symposium on Point-Based Graphics: 2007* Euro Graphics Society;91-100.
87. Epstein C, Carlsson G, Edelsbrunner H: **Topological data analysis.** *Inverse Probl* 2011, **27(12)**.
88. Shannon CE, Weaver W: **The Mathematical Theory of Communication.** Urbana (IL): University of Illinois Press; 1949.
89. Jeanquartier F, Holzinger A: **On Visual Analytics And Evaluation In Cell Physiology: A Case Study.** In *Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127.*. Heidelberg, Berlin: Springer;Cuzzocrea A, Kittl C, Simos DE, Weippl E, Xu L 2013:495-502.
90. Keim D: **Pixel-oriented visualization techniques for exploring very large databases.** *Journal of Computational and Graphical Statistics* 1996, **5(1)**:58-77.
91. Fayyad U, Grinstein GG, Wierse A: **Information Visualization in Data Mining and Knowledge Discovery.** San Francisco et al.: Morgan Kaufmann; 2002.
92. Kosara R, Miksch S: **Visualization methods for data analysis and planning in medical applications.** *International Journal of Medical Informatics* 2002, **68(1-3)**:141-153.
93. Lee JP, Carr D, Grinstein G, Kinney J, Saffer J: **The Next Frontier for Bio- and Cheminformatics Visualization.** *IEEE Computer Graphics and Applications* 2002, **22(5)**:6-11.
94. Ware C: **Information Visualization: Perception for Design (Interactive Technologies) 2nd Edition.** San Francisco: Morgan Kaufmann; 2004.
95. Inselberg A: **Visualization of concept formation and learning.** *Kybernetes: The International Journal of Systems and Cybernetics* 2005, **34(1/2)**:151-166.
96. Hauser H, Hagen H, Theisel H: **Topology-based methods in visualization.** London: Springer; 2007.
97. Wiltgen M, Holzinger A, Tilz GP: **Interactive Analysis and Visualization of Macromolecular Interfaces Between Proteins.** In *HCI and Usability for Medicine and Health Care Lecture Notes in Computer Science (LNCS 4799).*. Berlin, Heidelberg, New York: Springer;Holzinger A 2007:199-212.
98. Gehlenborg N, Brazma A: **Visualization of large microarray experiments with space maps.** *BMC Bioinformatics* 2009, **10(Suppl 13)**:O7.
99. Aigner W, Miksch S, Schumann H, Tominski C: **Visualization of Time-Oriented Data. Human-Computer Interaction Series.** London: Springer; 2011.
100. Pascucci V, Tricoche X, Hagen H, Tierny J: **Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications.** Berlin, Heidelberg: Springer; 2011.
101. Schroeder D, Kowalewski T, White L, Carlis J, Santos E, Sweet R, Lendvay TS, Reihsen T, Keefe DF: **Exploratory Visualization of Surgical Training Databases for Improving Skill Acquisition.** *Computer Graphics and Applications, IEEE* 2012, **32(6)**:71-81.

102. Wong BLW, Xu K, Holzinger A: **Interactive Visualization for Information Analysis in Medical Diagnosis.** In *Information Quality in e-Health, Lecture Notes in Computer Science, LNCS 7058.*. Springer Berlin Heidelberg;Holzinger A, Simonic K-M 2011:109-120.
103. Gigerenzer G: **Gut Feelings: Short Cuts to Better Decision Making.** London: Penguin; 2008.
104. Gigerenzer G, Gaissmaier W: **Heuristic Decision Making.** In *Annual Review of Psychology, Vol 62* Fiske ST, Schacter DL, Taylor SE 2011, 451-482, vol. 62. Palo Alto: Annual Reviews.
105. Tory M, Möller T: **Human Factors in Visualization Research.** *IEEE Transactions on Visualization and Computer Graphics* 2004, **10(1)**:72-84.
106. Munzner T, Johnson C, Moorhead R, Pfister H, Rheingans P, Yoo TS: **NIH-NSF visualization research challenges report summary.** *Computer Graphics and Applications, IEEE* 2006, **26(2)**:20-24.
107. Saad A, Hamarneh G, Möller T: **Exploration and Visualization of Segmentation Uncertainty using Shape and Appearance Prior Information.** *IEEE Transactions on Visualization and Computer Graphics* 2010, **16(6)**:1365-1374.
108. Weippl E, Holzinger A, Tjoa AM: **Security aspects of ubiquitous computing in health care.** *Springer Elektrotechnik & Informationstechnik, e&i* 2006, **123(4)**:156-162.
109. Holzinger A: **Successful Management of Research and Development.** Norderstedt: BoD; 2011.
110. Furniss D, O'Kane AA, Randell R, Taneva S, Mentis H, Blandford A: **Fieldwork for Healthcare: Case Studies Investigating Human Factors in Computing Systems.** *Synthesis Lectures on Assistive, Rehabilitative, and Health-Preserving Technologies* 2014, **2(1)**:1-129.
111. Berka P, Rauch J, Tomecková M: **Lessons Learned from the ECML/PKDD Discovery Challenge on the Atherosclerosis Risk Factors Data.** *Computing and Informatics* 2007, **26(3)**:329-344.
112. Fawcett T: **An introduction to ROC analysis.** *Pattern Recognit Lett* 2006, **27(8)**:861-874.
113. Holzinger A, Geierhofer R, Modritscher F, Tatzl R: **Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses.** *J Univers Comput Sci* 2008, **14(22)**:3781-3795.
114. Beale R: **Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and Web browsing.** *International Journal of Human-Computer Studies* 2007, **65(5)**:421-433.
115. Holzinger A, Kickmeier-Rust MD, Wassertheurer S, Hessinger M: **Learning performance with interactive simulations in medical education: Lessons learned from results of learning complex physiological models with the HAEMOdynamics SIMulator.** *Computers & Education* 2009, **52(2)**:292-301.
116. Allen B, Bresnahan J, Childers L, Foster I, Kandaswamy G, Kettimuthu R, Kordas J, Link M, Martin S, Pickett K, *et al*: **Software as a Service for Data Scientists.** *Communications of the ACM* 2012, **55(2)**:81-88.
117. Garg V, Arora S, Gupta C: **Cloud Computing Approaches to Accelerate Drug Discovery Value Chain.** *Combinatorial Chemistry & High Throughput Screening* 2011, **14(10)**:861-871.
118. Kagadis GC, Kloukinas C, Moore K, Philbin J, Papadimitroulas P, Alexakos C, Nagy PG, Visvikis D, Hendee WR: **Cloud computing in medical imaging.** *Med Phys* 2013, **40(7)**.
119. Lin CW, Abdul SS, Clinciu DL, Scholl J, Jin XD, Lu HF, Chen SS, Iqbal U, Heineck MJ, Li YC: **Empowering village doctors and enhancing rural healthcare using cloud computing in a rural area of mainland China.** *Computer Methods and Programs in Biomedicine* 2014, **113(2)**:585-592.
120. Kotseruba Y, Cumbaa CA, Jurisica I, Iop: **High-throughput protein crystallization on the World Community Grid and the GPU.** *High Performance Computing Symposium 2011* 2012, **341.**
121. Kreuzthaler M, Bloice MD, Simonic K-M, Holzinger A: **On the Need for Open-Source Ground Truths for Medical Information Retrieval Systems.** In *I-KNOW 2010, 10th International Conference on Knowledge Management and Knowledge Technologies; Graz (Austria)* Tochtermann K, Maurer H 2010, 371-381.
122. Kreuzthaler M, Bloice MD, Faulstich L, Simonic KM, Holzinger A: **A Comparison of Different Retrieval Strategies Working on Medical Free Texts.** *J Univers Comput Sci* 2011, **17(7)**:1109-1133.
123. Williams AJ, Ekins S, Tkachenko V: **Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation.** *Drug Discovery Today* 2012, **17(13-14)**:685-701.
124. Hall MA, Holmes G: **Benchmarking attribute selection techniques for discrete class data mining.** *Ieee Transactions on Knowledge and Data Engineering* 2003, **15(6)**:1437-1447.
125. Holzinger A, Yildirim P, Geier M, Simonic K-M: **Quality-Based Knowledge Discovery from Medical Text on the Web.** In *Quality Issues in the Management of Web Information, Intelligent Systems Reference Library, ISRL 50.*. Berlin Heidelberg: Springer;Pasi G, Bordogna G, Jain LC 2013:145-158.
126. Begley CG, Ellis LM: **Drug development: Raise standards for preclinical cancer research.** *Nature* 2012, **483(7391)**:531-533.
127. Patrello C, Pasini E, Kotlyar M, Otasek D, Wong S, Sangrar W, Rahmati S, I. J: **Integration, visualization and analysis of human interactome.** *Biochemical and Biophysical Research Communications* in press; 2014.
128. Ponzielli R, Boutros PC, Katz S, Stojanova A, Hanley AP, Khosravi F, Bros C, Jurisica I, Penn LZ: **Optimization of experimental design parameters for high-throughput chromatin immunoprecipitation studies.** *Nucleic acids research* 2008, **36(21)**:e144-e144.

doi:10.1186/1471-2105-15-S6-I1
Cite this article as: Holzinger *et al.*: Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. *BMC Bioinformatics* 2014 **15**(Suppl 6):I1.